International Journal of Knowledge *www.ijklp.org*
and Language Processing KLP International ⓒ2013 ISSN 2191-2734
Volume 4, Number 3, **2013** pp.**17--26**

Medical literature's semantic status and solution

Zhang Jianping and Liu Yao

Institute of Scientific and Technical Information of China
No.15, Fuxing Road, Haidian District, Beijing, China
zjping51@126.com
liuy@istic.ac.cn

Received June 2013; revised August 2013

ABSTRACT. *This article presents the status of literature's semantic, especially the medical literature, and the insufficient of semantic annotation tools. Facing with those problems, we pretend to use NLP techniques and machine learning methods to construct skin disease ontology automatically to solve these problems.*

Keywords: Semantic Annotation, Medical Literature, Natural Language Processing

1. **Introduction.** Web founder Tim Berners-Lee put forward the concept of the Semantic Web in 1998 , with the W3C officially launched Semantic Web Activity in February 2001, semantic retrieval became the mainstream of network research and development. Facing of the geometric published biomedical literature (BIO), how to reveal the semantic information from the text data is becoming important. Semantic, is to choose the appropriate semantic tags, with what semantic features reflected, so that information becomes a machine-processable information. There are two key issues involved in Semantics: the generation of semantic metadata and semantic annotation.

Semantic metadata (also known as Tag Ontology) provides the data's semantic information, it is also an important tool for organizing library semantic resources. Semantic metadata's generation technology and domain knowledge base or ontology's construction, from the technology point of view, they are consistent. Its technology is quite mature. Hanxian Pei[1] et al. put forward a Wikipedia-based construction method, the target semantic metadata was extracted by analyzing the table of contents entries in a category (table-of-contents), and according to the title of the section of the document to give the instance a semantic markup correctly, and ultimately build a training corpus. Solve two problems in building automatic semantic metadata's generation system: target semantic metadata extraction and corpus construction, laid the foundation for further large-scale applications of semantic metadata. Liu Yao et al[2] proposed the use of natural language processing (NLP) theory and technology has been generally accepted in the field of knowledge, such as professional thesaurus, professional dictionary, professional teaching or authoritative book reconstruction and use; with the help of experts in the field of knowledge, web-based knowledge acquisition and processing; establish a limited text ontology self-learning mechanism, in order to achieve field ontology concept describe the system automatically constructed.

Semantic annotation is marking text or symbols of the original data, to have semantic information, not only the man can understand, but also the machine can understand. Semantic Web is generally in XML markup language data for label , described as a data model in RDF / XML, and combining Ontology, the data is marked with a clear meaning, so that the machine can understand. Semantic annotating tool today has been formed there such as Gate ， OntoMat Annotizer, Seeker, MnM SHOE Knowledge Annotator, Annotea SMORE SemanticWord, Melita, SemanticMarkup Plug-In Briefing Associate Yawas Annozilla 10 [3], they are all ontology-based annotating tool. Now, there are normaly three Semantic annotation methods[4] : (1) Artificial annotation, by experts to determine the applicable set of concepts, metadata elements, the establishment of the semantic data in RDF or HTML language tags. This process requires a lot of manpower and material resources. (2) the use of Document Type Definition (DTD) and document mode (Schema) concept mapping and annotating, because of the document’s structure and element was definited detailed by the SGML/XML document, so that after the establishment of specific concepts’ special mapping of the DTD/Schema, the SGML/XML document’s DTD/Schema’s content element tag can be transformed to homologous concept element tag automaticly, but this method also needs to be checked and modified manually; (3)annotation with analysis of the vocabulary’s meaning ,the method mainly uses automaticly analysis and extracting the vocabulary to establish the mapping of the vocabulary and the concept’s classification, then find the document’s concept classification, even its annotation relationship with other class, to annotate with these concepts. But this method has domain restriction and also needs to be checked manually. Now, there are more mature Semantic annotating Platform such as AeroDAML, MnM, MUSE, Armadillo, KIM, SemTag, Ont-O-Mat, etc.

2. **Problem Statement and Preliminaries.** To reducing the man’s participate in the generation of semantic metedata, some semi-automatic assistance skills were used to these annotating tools. In the part of constructing system, CREAM system’s expanding researches of S_CREAM[5] and PASNKOW[6], S_CREAM achieved semi-automaticly annotation of the web, which is based on Amilcare’s information extraction. PANKOW （Pattern-based Annotation through Knowledge on the Web） uses a studying method based on unsupervised framework to classify the example to ontology, at first processes the web page, find nouns may be became ontology ,then accord to a schema to generate phrases contained in the noun, using Google API to find the times of the phrase, use the calculated weight to determine the noun’s classification in the ontology system. To be annotated automatically, a lot of work focus on designing semantic annotation metadata ‘s generate automatically model. Hung[7]et al. Based on Web resource mining, acquires training corpora necessary to describe both the thematic categories and the metadata

extracted from the texts. The approach then finds the corresponding relationships among them by means of categorization and thus generates thematic metadata for the textual data. Yang and Lee[8]proposed a method using machine-learning to generate web page's semantic metedata automatically. The proposed automated process adopts the self-organizing map algorithm to cluster training Web pages and conducts a text mining process to discover some semantic descriptions about the Web pages. A.Dingli[9]proposed a framework of Armadillo. The methodology is based on a combination of in-formation extraction, information integration and machine learning techniques. Learning is seeded by extracting information from structured sources (e.g. databases and digital libraries). Retrieved information is then used to partially annotate documents. These annotated documents are used to bootstrap learning for simple Information Extraction (IE) methodologies, which in turn will produce more annotations used to annotate more documents. H.Graubitz[10] displayed DIAsDEM framework , it is based on cluster text units into homogeneous groups whose labels form the XML tags surrounding their contents and from which the document type definition for the collection is derived. J.Li[11]described a mechine-learing method, it's semantic annotation is based on sentences, and converted to RDF, it adopt a dependency grammar – Link Grammar for this purpose Dill[12]describes Seeker, a platform for large-scale text analytics, and SemTag, an application written on the platform to perform automated semantic tagging of large corpora. We apply SemTag to a collection of approximately 264 million web pages, and generate approximately 434 million automatically disambiguated semantic tags, published to the web as a label bureau providing metadata regarding the 434 million annotations.

Medical literature's semantic metadata's generation is similar of metadata. Liu yao[13]proposed a construction method of Chinese medicine ontology based on history literature, use of history literature and authority domain knowledge to illustrate Chinese medicine's core concept. To construct and obtain Chinese medical knowledge system automatically. Meeraman ,R[14]proposed to construct ontology based on NLP, this method uses WWW to analysis domain literature, can ignore linguistic knowledge and resource ,suggest a linguistic friendly interface to enhance language processing based on ontology such as semantic annotation. Deng ziping[15]proposed some methods to construct domain ontology manually. With the source of electronic medical record for research, focus on solving three key question of construction of medical domain ontology, research on medical diagnosis and treat ontology system's automatically generate theory and technology. Fang an[16]and Fu qiang[17]' research is based on Hadzic of Australian who proposed four-dimensional medical ontology model , improve it to construct HFMD (Hand, Foot and Mouth Disease,)and arrhythmia ontology. Mo dongmei[18]discussed the necessity of thesaurus in the construction of medical ontology, and the construction flow of prevent medicine ontology, including ontology construction tool, method, and the convert from thesaurus to domain ontology, using racer inference machine to inference domain ontology.

Zhou xiaojia[19]proposed a solution of automatic extraction of temporal attributes of medical problems from Chinese narrative medical records based on conditional random fields (CRF). Gao yonggang[20]used MI and its improved MI greedy optimization algorithm to achieve medical image feature selection; implement MI-based feature selection with support vector machine (SVMs) classification labels, so as to establish low-level visual features of medical images and high-level semantic features of the mapping between; solve medical image annotation in the "semantic gap" problem and all the characteristics of large amount of calculation involved in classification problems. Mi Yang[21]'s research is based on top ontology' integration of medical semantic annotation, it is of ontology's construction, ontology mapping and integration, semantic annotation, etc., explore the medical field relational structure of the top ontology semantic ontology semantic annotation working mechanism, try to establish a top-level ontology integration based on the medical domain ontology semantic annotation system model, empirical semantic annotation new ideas, explore ontology integration effect for semantic annotation role in promoting and strive to achieve from reality, to study and solve practical application of existing ontology semantic barriers, thereby improving medical information resources level semantic annotation. Tsatsaronis, George[22]proposed a method based on a Maximum Entropy approach, for annotating biomedical literature documents with terms from the Medical Subject Headings (MeSH).The experimental evaluation shows that the suggested Maximum Entropy approach for annotating biomedical documents with MeSH terms is highly accurate, robust to the ambiguity of terms, and can provide very good performance even when a very small number of training documents is used. Ruzicka, M; Svatek, V[23] proposed an annotation system—stepper, the system can extract information from literatures that are unique, authority, and rich in ambiguously expressed knowledge. Yu jiang de[24]proposed a method based on Conditional Random semantic role labeling, the method is based on shallow parsing, phrases or named entities is marked as the basic unit of the semantic annotation. CRFs model is used to predicate sentence semantic role labeling, experimental results show that: based on Conditional Random approach than a method based on maximum entropy better performance of the method in the semantic role labeling task gained 80.43% and 63.55% accuracy rate of recall. Volk, M[25]present a framework for concept-based cross-language information retrieval in the medical domain, which is under development in the MUCHMORE project.

Current ontology-based semantic annotation tools all have their own characteristics and scope, but the existing tools of the following deficiencies: 1. No tools support the latest W3C ontology language OWL; 2.most tools do not support ontology vocabulary expansion; 3.majority tools do not support simultaneous open, browse multiple ontology and ontology annotation using multiple pages; 4.mark all the tools only in English, does not support multi-language; 5.label objects tools is static content-based; 6. most of the tools used first create the content, after marked "two-step", only a few tools support content writing and semantic annotation simultaneously; 7. semantic annotation process of ontology query, assisted reasoning support, and the degree of automation of metadata generated is not enough [26].

3. The theoretical basis of the Experiment. To solve these problems, Liu Yao [27], etc.

According to the theory and bibliographic control conjugate theory using natural language processing (NLP) techniques and machine learning methods have been accepted domain knowledge, such as professional thesaurus, professional dictionaries, textbooks or professional the use of such authoritative works to reconstruct, build domain ontology, based on the development of Chinese literature resources semantic annotation technology for semantic annotation of the relevant literature, and in the relatively large number of semantic content, based on the combination of traditional organizational resources (thesauri, etc.), through machine learning and other methods to generate the initial semantic metadata, and then with the help of the auxiliary platform library resources to achieve organizational processes and semantic semantic metadata system construction simultaneously, and semantic annotation semantic indexing literature and stored separately.

In this theory, the authors will take the skin diseases for example, where the Mesh terms in the auxiliary platform Ontology as the knowledge elements, its classification and property settings to be adjusted and modified in order to build the knowledge described in the prototype system. Prepared related literature, import the teaching materials into the system, to establish skin disease ontology automaticly, to verify and improve the method.

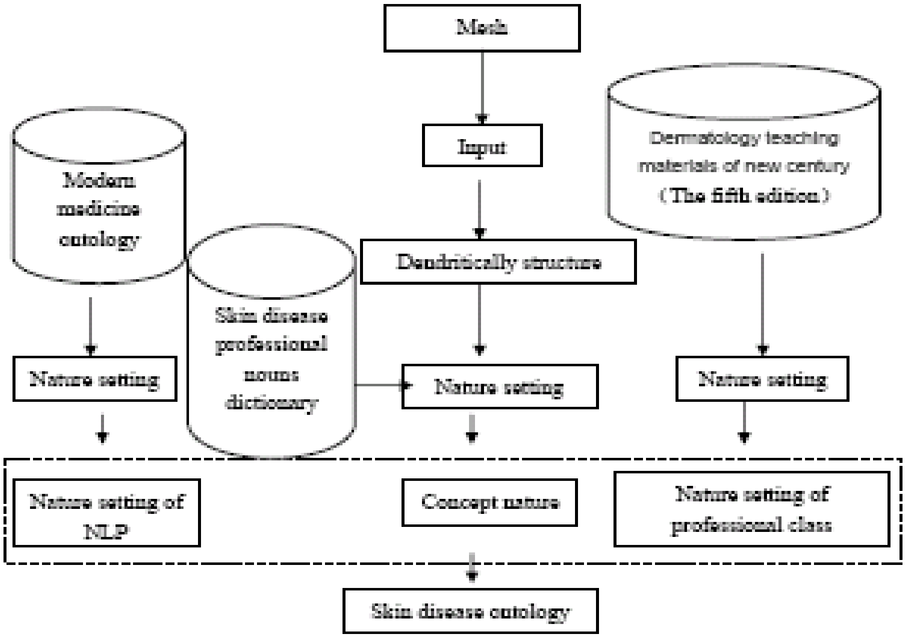
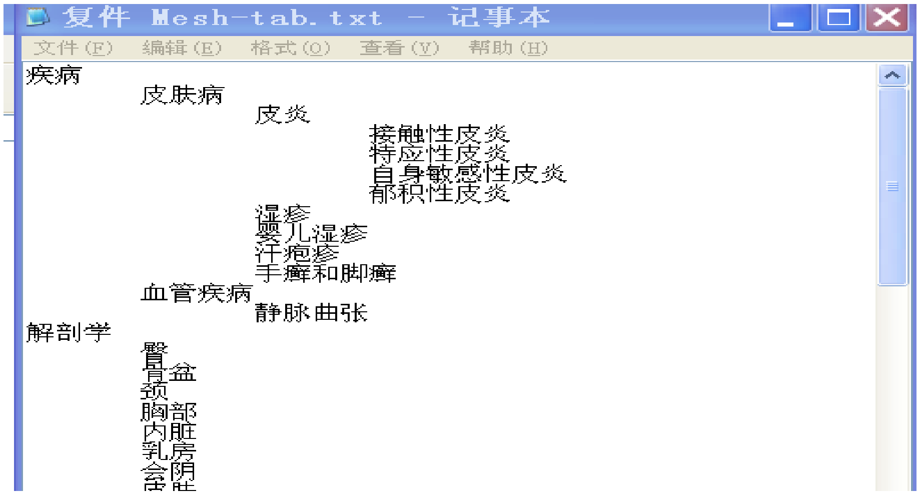


FIG.1.THE METHOD'S CONSTRUCTION SCHEDULE

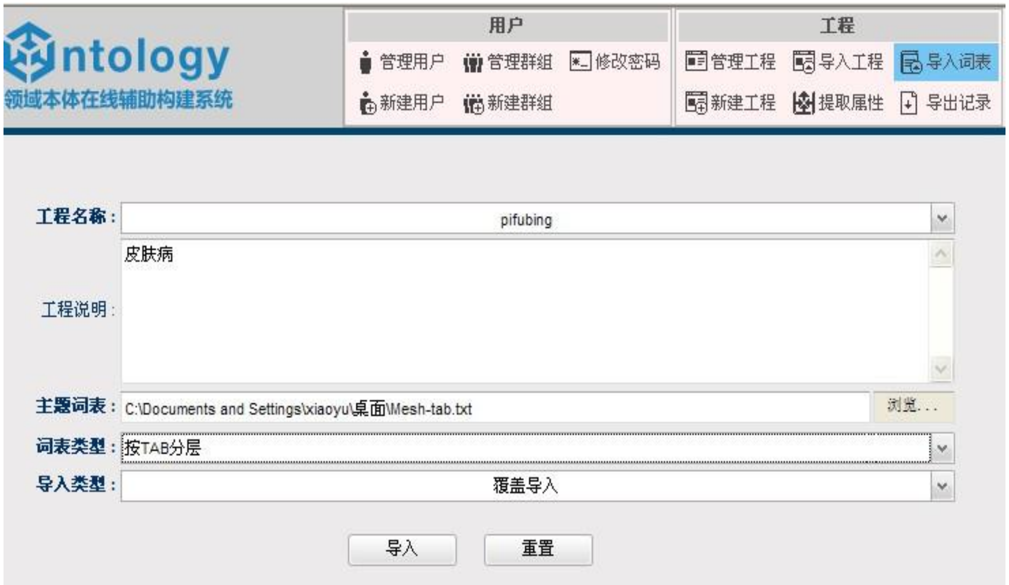
4. **The experiment's key technologie.**

4.1. **Input of lexicon.** The lexicon we selected was Medical Subject Headings, which is editioned by NLM , we selected its part of skin disease and part that has relationships with the disease. This lexicon is structured, covers almost all skin diseases. We download the lexicon; sort it to unique format which the system asked for. The format of their hierarchical relationships, up and down relationship is represented with the Tab key hierarchy, and stored in txt. of UTF-8 format.



```
疾病
            皮肤病
                        皮炎
                                    接触性皮炎
                                    特应性皮炎
                                    自身敏感性皮炎
                                    郁积性皮炎
                        湿疹
                        婴儿湿疹
                        汗疱疹
                        手癣和脚癣
            血管疾病
                        静脉曲张
解剖学
            臀
            骨盆
            颈
            胸部
            内脏
            乳房
            会阴
```

FIG.2.LEXCION THAT THE SYSTEM REQUIRES



工程名称：pifubing

工程说明：皮肤病

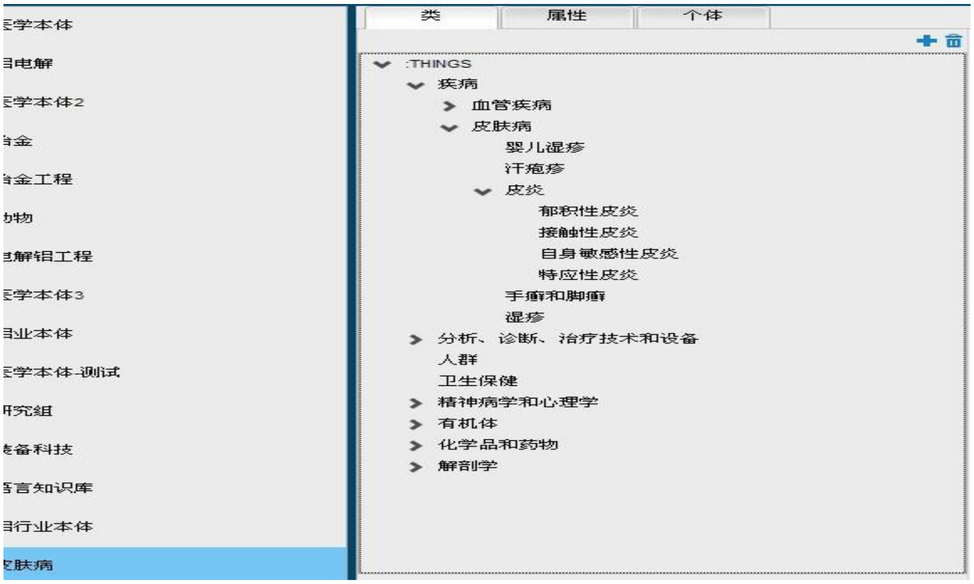
主题词表：C:\Documents and Settings\xiaoyu\桌面\Mesh-tab.txt

词表类型：按TAB分层

导入类型：覆盖导入

FIG.3.INPUT THE LEXICON TO THE SYSTEM

Input this lexicon to the platform， and after inputting the lexicon, the platform generates the class of dendritically structure as follows:

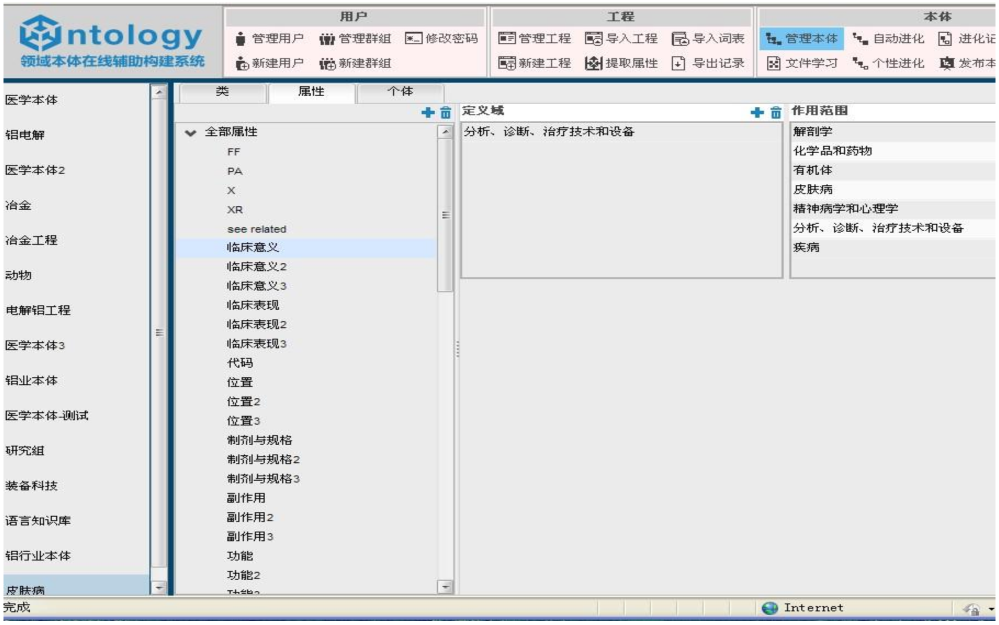


学本体
电解
学本体2
金
金工程
物
解铝工程
学本体3
业本体
学本体-测试
究組
备科技
言知识库
行业本体
肤病

- :THINGS
  - 疾病
    - 血管疾病
    - 皮肤病
      - 婴儿湿疹
      - 汗疱疹
      - 皮炎
        - 郁积性皮炎
        - 接触性皮炎
        - 自身敏感性皮炎
        - 特应性皮炎
      - 手癣和脚癣
      - 湿疹
  - 分析、诊断、治疗技术和设备
  - 人群
  - 卫生保健
  - 精神病学和心理学
  - 有机体
  - 化学品和药物
  - 解剖学

FIG.4. DENDRITICALLY STRUCTURE OF THE PLATFORM

4.2. **Nature setting.** Nature setting contains three parts .Part one is nature setting of the class, every class has public nature ,also has its unique class, for example disease has its public class FF, PA, X, XR, see related, code, also has its nature of Etiology and pathogenesis、clinical manifestations、diagnosis、treatment. So as the other class. after all classes ‘s nature has been seted , Part two is the setiing of the nature’s domain of definition, every nature should be setted , Part three is the setting of the nature’s actuating range, all the nature and nature with the number 2 should be setted.



医学本体
铝电解
医学本体2
冶金
冶金工程
动物
电解铝工程
医学本体3
铝业本体
医学本体-测试
研究組
装备科技
语言知识库
铝行业本体
皮肤病

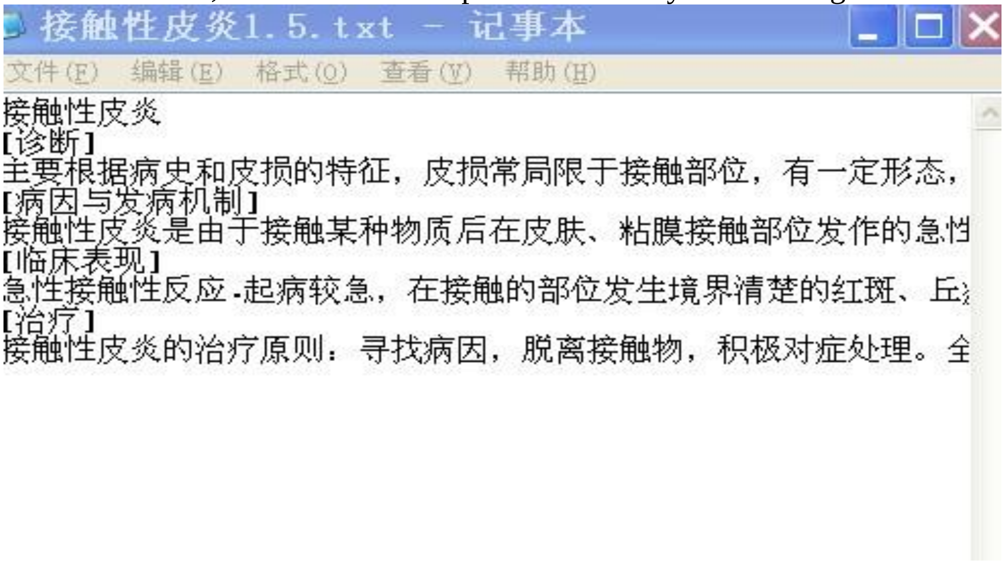
- 全部属性
  - FF
  - PA
  - X
  - XR
  - see related
  - 临床意义
  - 临床意义2
  - 临床意义3
  - 临床表现
  - 临床表现2
  - 临床表现3
  - 代码
  - 位置
  - 位置2
  - 位置3
  - 制剂与规格
  - 制剂与规格2
  - 制剂与规格3
  - 副作用
  - 副作用2
  - 副作用3
  - 功能
  - 功能2
  - 功能3

定义域: 分析、诊断、治疗技术和设备

作用范围:
解剖学
化学品和药物
有机体
皮肤病
精神病学和心理学
分析、诊断、治疗技术和设备
疾病

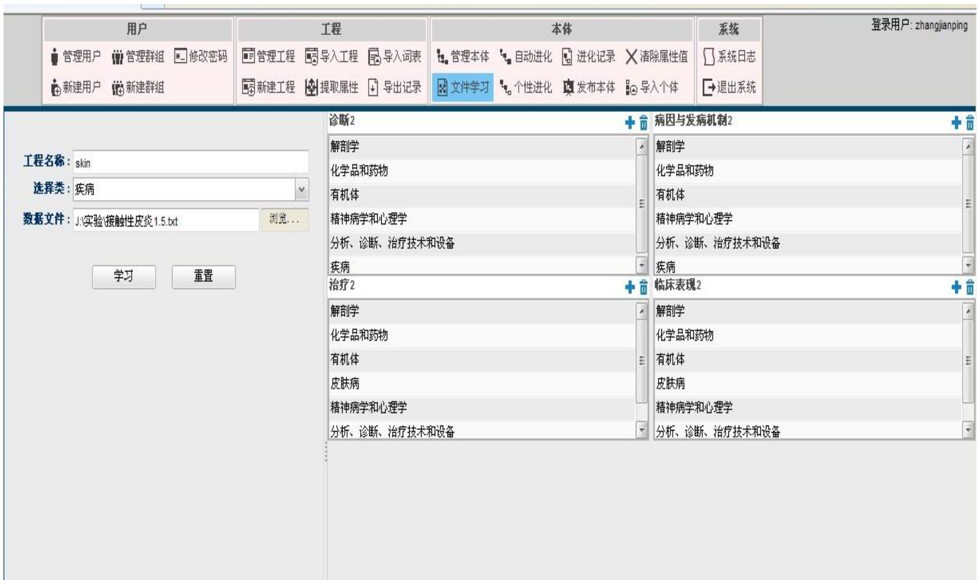
FIG.5.NATURE SETTING OF THE CLASS

4.3. **Format settings.** After the setting of the nature, to input the studying material, the studying material format should be setted to the format that the system requires, then the platform can identifies ,so that can use the platform to study the teaching materials.



接触性皮炎
[诊断]
主要根据病史和皮损的特征，皮损常局限于接触部位，有一定形态，
[病因与发病机制]
接触性皮炎是由于接触某种物质后在皮肤、粘膜接触部位发作的急性
[临床表现]
急性接触性反应-起病较急，在接触的部位发生境界清楚的红斑、丘
[治疗]
接触性皮炎的治疗原则：寻找病因，脱离接触物，积极对症处理。全

FIG.6.STUDING MATERIAL’ FORMAT



登录用户: zhangjianping

工程名称: skin
选择类: 疾病
数据文件: J:\实验\接触性皮炎1.5.txt

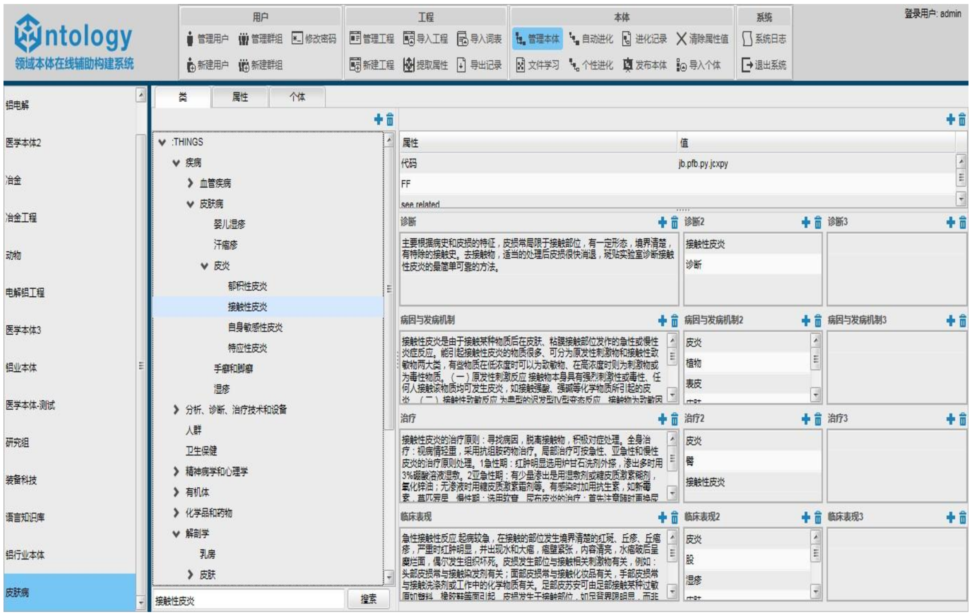
诊断2
解剖学
化学品和药物
有机体
精神病学和心理学
分析、诊断、治疗技术和设备
疾病

病因与发病机制2
解剖学
化学品和药物
有机体
精神病学和心理学
分析、诊断、治疗技术和设备
疾病

治疗2
解剖学
化学品和药物
有机体
皮肤病
精神病学和心理学
分析、诊断、治疗技术和设备

临床表现2
解剖学
化学品和药物
有机体
皮肤病
精神病学和心理学
分析、诊断、治疗技术和设备

FIG.7.STUDYING METHOD



接触性皮炎

属性	值
代码	jb.pfb.py.jcxpy
FF	
see related	

诊断

主要根据病史和皮损的特征，皮损常局限于接触部位，有一定形态，境界清楚，有特除的接触史。去接触物，适当的处理后皮损很快消退，斑贴实验室诊断接触性皮炎的最简单可靠的方法。

诊断2

接触性皮炎
诊断

病因与发病机制

接触性皮炎是由于接触某种物质后在皮肤、粘膜接触部位发作的急性或慢性炎症反应。能引起接触性皮炎的物质很多、可分为原发性刺激物和接触性致敏物两大类，有些物质在低浓度时可以为致敏物、在高浓度时则为刺激物或为毒性物质。（一）原发性刺激反应 接触物本身具有强烈刺激性或毒性、任何人接触该物质均可发生皮炎，如接触强酸、强碱等化学物质所引起的皮炎 （二）接触性致敏反应 为典型的迟发型IV型变态反应，接触物为致敏因

病因与发病机制2

皮炎
植物
表皮

治疗

接触性皮炎的治疗原则：寻找病因，脱离接触物，积极对症处理。全身治疗：视病情轻重，采用抗组胺药物治疗。局部治疗可按急性、亚急性和慢性皮炎的治疗原则处理。1急性期：红肿明显选用炉甘石洗剂外搽，渗出多时用3%硼酸溶液湿敷。2亚急性期：有少量渗出是用湿敷剂或糖皮质激素糊剂，氧化锌油；无渗液时用糖皮质激素霜剂等。有感染时加用抗生素，如新霉素，莫匹罗星 慢性期：选用软膏 尿布皮炎的治疗：首先注意随时更换尿

治疗2

皮炎
臀
接触性皮炎

临床表现

急性接触性反应.起病较急，在接触的部位发生境界清楚的红斑、丘疹、丘疱疹，严重时红肿明显，并出现水和大疱，疱壁紧张，内容清亮，水疱破后呈糜烂面，偶尔发生组织坏死。皮损发生部位与接触相关刺激物有关，例如：头部皮损常与接触染发剂有关；面部皮损常与接触化妆品有关，手部皮损常与接触洗涤剂或工作中的化学物质有关。足部皮苏安可由足部接触某种过敏原如塑料、橡胶鞋等而引起，皮损发生于接触部位，如足背界限明显，而非

临床表现2

皮炎
股
湿疹

FIG.8.STUDYING RESULT

5. **Conclusions.** The above is the studying of one material , it also can study other material , to Realize the function. This is about skin disease ontology automated build and annotation system development, this platform also need to further improve the indexing efficiency and accuracy.

Acknowledgment. This work is partially supported by National Social Science Fund No. 12BTQ006, National Key Project of Scientific and Technical Supporting Programs No. 2013BAH21B02; the authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] HAN Xianpei, ZHAO Jun. Semantic Metadata Generation: A Method Based on Wikipedia［J］. Journal of Chinese Information Processing, 2009, 23 (2) : 108-114.

[2] LIU Yao , SUI Zhi fang , HU Yong wei, et al. Automatic construction of domain ontology [J]. Journal of Beijing University of Posts and Telecommunications, 2006, 29 (Z2) : 65- 69.

[3] http://semanticweb.org/wiki/Tools

[4] Zhang Xiaolin. Semantic Web and Semantic-based Networked Information Retrieval, Journal of The China Society For Scientific and Technical Information ,2002, 14 (21).

[5] HANDSCHUH S, et al. S-CREAM-Semi-automatic CREAtion of metadat［C］ 13th International Conference, EKAW 2002 Sigüenza, Spain, 2002: 358-372.

[6] CIMIANO P, et al. Towards the self annotating Web[C] FELDMAN S I, et al. Proceedings of WWW,2004:462-471.

[7] CHIE CHUNG HUANG, et al. Using a Web based categorization approach to generate thematic

metadata from texts[J]. ACM Transactions on Asian Language Information Processing，2004，3 (3) : 190-212.

[8] YANG H-C，LEE C-H. Automatic metadata generation for Web pages using a text mining approach ［C］ International Workshop on Challenges in Web Information Retrieva1 and Integration，2005: 186-194.

[9] DINGLI A， et al. Automatic semantic annotation using unsupervised information extract on and integration ［C］ GENNARIJ，et al. Proceedings K-CAP，2003.

[10] GRAUBITZ H, et al. Semantic tagging of domain-specific text documents with diasdem［C］SAAKE G， et al. Proceedings of DBFusion 2001. USA: ACM，2001: 61-72.

[11] LI J，et al. Learning to generate semantic annotation for domain specific sentences ［C］ GIL Y， et a1. Proceedings of KCAP，2001: 44-57.

[12] DILL S，et al. A case for automated largescale semantic annotation［J］. Web Semantics: Science，Services and Agents onthe World Wide Web，2003，1 (1) : 115-132.

[13] Liu Yao, Sui Zhifang, Zhou Yang .et al.Research on Automatic Construction of Chinese Traditional Medicine Ontology Concept' s Description Architecture. New Technology of Library and Information Service . 2008:5

[14] Meersman, R Creating ontologies for content representation - The OntoSeed suite

[15] Deng zi ping. Research and Development of a Ontology Automatic Generation System Oriented Medical Diagnosis. GuangDong University Of Technology[D]

[16] Fang an et al.Method Research of Constructing Clinical Disease Domain Ontology——A Case Study of Hand Foot and Mouth Disease Ontology. JOURNAL OF INTELLIGENCE.2009.v.28.No.11

[17] Fu qiang。Research on Medical Domain Ontology Building Based on Thesaurus. Jilin University.[D]

[18] Mu Dongmei ,Fan Yi .Building and Reasoning of Digital Library's Domain Ontology:Taking Medicine Domain Ontology as an Example. Library and Information Service.[J]. Vol.51,No.8,August,2007

[19] ZHOU Xiao-Jia LI Hao-Min DUAN Hui-Long .et al.The Automatic Extraction of Temporal Relation from Chinese Narrative Medical Records Using Conditional Random Fields. Chinese Journal of Biomedical Engineering[J]. Vol. 29 No. 5, 2010

[20] Gao yonggang.Research and Application on Semantic Annotation Technologies of Medical Images[D]. Northwestern University. 2009

[21] Mi yang. Research on Semantic Annotation Based on Upper-level Ontology Integration in Medical Field.[D].JiLin university

[22] Tsatsaronis, George. A Maximum-Entropy approach for accurate document annotation in the biomedical domain.

[23] Ruzicka, M; Svatek, V .Stepper: Annotation and interactive stepwise transformation for knowledge-rich documents[C]. 14th International Conference on Knowledge Engineering and Knowledge Management

[24] Yu Jiangde,Fan Xiaozhong1 Pang Wenbo1. Semantic role labeling based on conditional random fields. Journal of Southeast University(English Edition.[J]

[25] Volk, M. Semantic annotation for concept-based cross-language medical information retrieval

[26] TAO Wan, LI Ping, LIAO Shu-mei. Analysis and summary of current ontology-based semantic annotation tools Journal of Anhui University of Technology and Science, 2005 ,20(2)

[27] Liu Yao et al.，Research on the Method of Library Resources Organization Semantization Based on Content-format Interaction[J]. Information Studies:Theory & Application,. 2010 ,33(10)